IJDC | *General Article*

MOLES3: Implementing an ISO standards driven data catalogue

Graham A Parton, Steve Donegan, Stephen Pascoe, Ag Stephens, Spiros Ventouras
Centre for Environmental Data Archival
STFC Rutherford Appleton Laboratory
U.K.

Bryan N Lawrence
NCAS, Department of Meteorology
University of Reading
U.K.

Abstract

ISO19156 Observations and Measurements (O&M) provides a standardised framework for organising information about the collection of information about the environment. Here we describe the implementation of a specialisation of O&M for environmental data, the Metadata Objects for Linking Environmental Sciences (MOLES3).

MOLES3 provides support for organising information about data, and for user navigation around data holdings. The implementation described here, "CEDA-MOLES", also supports data management functions for the Centre for Environmental Data Archival, CEDA. The previous iteration of MOLES (MOLES2) saw active use over five years, being replaced by CEDA-MOLES in late 2014. During that period important lessons were learnt both about the information needed, as well as how to design and maintain the necessary information systems. In this paper we review the problems encountered in MOLES2; how and why CEDA-MOLES was developed and engineered; the migration of information holdings from MOLES2 to CEDA-MOLES; and, finally, provide an early assessment of MOLES3 (as implemented in CEDA-MOLES) and its limitations.

Key drivers for the MOLES3 development included the necessity for improved data provenance, for further structured information to support ISO19115 discovery metadata export (for EU INSPIRE compliance), and to provide appropriate fixed landing pages for Digital Object Identifiers (DOIs) in the presence of evolving datasets. Key lessons learned included the importance of minimising information structure in free text fields, and the necessity to support as much agility in the information infrastructure as possible without compromising on maintainability both by those using the systems internally and externally (e.g. citing in to the information infrastructure), and those responsible for the systems themselves. The migration itself needed to ensure continuity of service and traceability of archived assets.

Submitted 16 January 2015 | *Accepted* 10 February 2015

Correspondence should be addressed to Graham Parton, Centre for Environmental Data Archival, STFC Rutherford Appleton Laboratory, Didcot OX11 0QX. Email: graham.parton@stfc.ac.uk

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v10i1.365

Introduction

The UK Centre for Environmental Data Archival (CEDA) holds a wealth of data, with over 2 PB in 143 million unique files organised into 238 collections as of October 2014. The data are actively curated, a process involving managing and migrating metadata, and providing user services to discover, understand, download and manipulate the data. However, despite years of managed data ingestion, and actively promoting the use of standards with the data provider community, there are still many different data formats and a vast spectrum in the quality and type of internal metadata. Hence important services include the provision of catalogues, documentation, and indexes, so the collections are manageable, discoverable, understandable and accessible by both present and future users.

The relationship between content, catalogue and other types of metadata is complex. Lawrence et al. (2009) introduced a dataset metadata taxonomy which, amongst others, described four key kinds of metadata: "discovery metadata", which provides high level dataset metadata suitable for finding interesting data; "character metadata", which provides assertions about the data, typically in citations; "browse metadata", which provides documentation suitable for deeper understanding of data, or for choosing between similar datasets; and, "archive metadata", which describes key parameters of the data resource as well characteristics of the physical archiving.

Lawrence et al. also introduced a content standard suitable for environmental browse metadata: Metadata Objects for Linking Environmental Sciences, or MOLES, which has seen a number of iterations over the last decade (MOLES1, MOLES2 and finally MOLES3.4). In this paper the CEDA implementation MOLES2 is reviewed, leading to the reasons why MOLES3.4, a specialisation of ISO19156 Observation and Measurements (O&M), was then developed. Implementation issues with the CEDA version of MOLES3, CEDA-MOLES, are then presented, before the penultimate section discusses limitations of the implementation. The relationship between all these structures – browse metadata, archive metadata, discovery metadata, MOLES3, ISO19156 and CEDA-MOLES – is depicted in Figure 1.

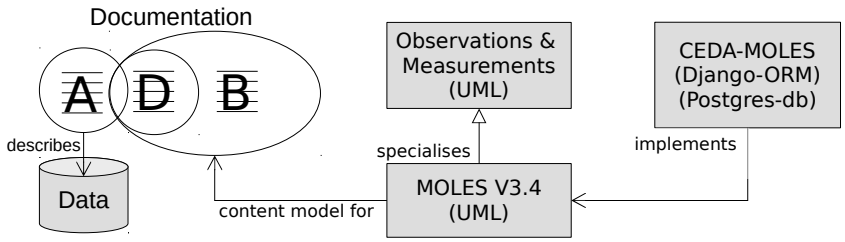

Figure 1. The relationship between information structures described in Lawrence et al. (2009), and the information structures described here. Browse metadata (**B**) is designed to support documentation and navigation around the information describing data holdings, as well as to support the export of content into discovery metadata. Browse metadata (and the discovery metadata – **D** – which it can export) may include information harvested from archive metadata (**A**). MOLES3.4 is a specialisation of ISO19156 Observations and Measurements (both of which are defined in the Unified Modelling Language, UML) which delivers a content model to support the Browse functionality, and it is implemented in CEDA-MOLES which exploits the Django object-relational mapper to manipulate metadata held in a Postgres database.

Typical data deployments often concentrate on a combination of the archive and discovery metadata, with the latter providing the content which underpins both institutional and federated portals, and the former being "in" downloadable data files. As Lawrence et al. noted, however, key steps in finding and eventually using data include understanding enough about the data to ascertain their fitness-for-purpose, and discriminating and navigating between similar offerings. It is that (organised) documentation, termed "browse" metadata, which may include material harvested from archive metadata, and may in turn be harvested for discovery.

The functional requirement for supporting navigation in browse metadata was one motivation for designing MOLES. The key concept of the linkages was to enable data consumers to find data via navigating around aspects of the data provenance. For example, having found a particular dataset produced by a specific instrument, find other datasets produced by the same instrument – without resorting to a high level (discovery based) search – and then discriminate between these datasets based on other, more detailed aspects of the metadata. Comprehensive browse metadata in MOLES was also intended to serve as the basis from which discovery metadata could be extracted.

In early versions of MOLES, the main information classes were Data Entities, Deployments, Activities, Data Production Tools (DPTs), and Observation Stations (OBS). The relationship between these entities is shown in Figure 2. Within MOLES2, as implemented in CEDA, a Data Entity described *what* data are in the archive and was connected, via a relevant Deployment, to the reason *why* the data were collected, detailed by an Activity record; a description of *how* the data were produced, through one or more DPT records (so called to be agnostic between simulations and measurements); and an OBS record detailing *where* the data were produced.

Each entity was intended to describe a particular aspect of either the data itself or relevant background information, in such a way that where different aspects of the information were shared by different entities, their descriptions were implemented as shared entities. Such information reuse provided the route for users to navigate via all the associations from any entity, delivering the "browse" function described earlier, as well as background information to aid data provenance. In principle, shared Activity, DPT and OBS records would also provide ways to aggregate datasets into more complex assemblages with multiple Deployments.

In practice, various issues arose with the MOLES2 implementation. Firstly, a number of key attributes, such as named parties, were not reusable leading to duplicate, unlinked and often inconsistent entries both within and between records. Secondly, without constraints on use, record administrators subverted the use of the different record types from their original intended purposes, for example by linking to the archive from any record type and not just from Data Entities or Deployments, leading to confusion with both administrators and end users. Thirdly, it lacked elements required for exporting content into downstream services. This latter issue could not be entirely resolved within MOLES2, forcing a reliance on a partial solution using bespoke mark-up within text fields of the MOLES2 records and ad-hoc manipulation of content within scripts during export to external discovery services. Such external demands for content came primarily from two directions: to support external ISO compliant metadata to meet the EU INSPIRE legislation[1], and the need to support the information content for data Publication

[1] http://inspire.ec.europa.eu/

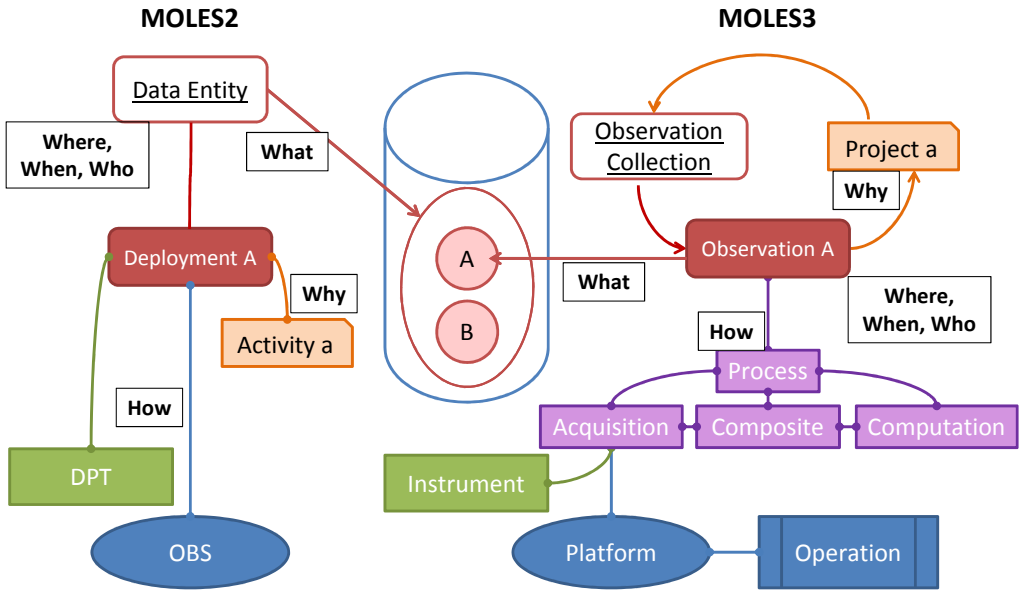

Figure 2. Schematic of main MOLES2 and MOLES3 entities. In MOLES2 (left hand side), Data entities (describing *what* is in the archive) are linked via Deployments to the reasons *why* the data were collected (Activities), *how* they were collected (Data Production Tools, DPTs) and *where* (Observation Stations, OBS). In MOLES3 (right hand side), broadly equivalent concepts (indicated by the use of similar colours and shapes) for *how, what, why* and *where* exist, but significantly more detail is expected in the *how*, via the Process description and the Acquisition, Computation and Composite classes (where Composite represents aggregations of Acquisition and Computation). In both cases, the people (or parties) involved, and recorded as entity attributes, should also provide a route for direct navigation between entities.

(e.g. formal citations for Digital Object Identifiers [DOIs] and their concomitant landing pages [Callaghan et al., 2012]).

MOLES 3: A New Content Standard

Given issues in the MOLES2 information model, its implementation and the increasing workflow demands for exporting discovery metadata to external services, the obvious solution was to evolve the MOLES standard and revise the implementation methodology.

In practice the evolution involved a number of steps: Firstly, the standards environment was revisited to see if new or different opportunities for exploitation had evolved, and ISO19156: Observations and Measurements (O&M) (International Organization for Standardization, 2013) was chosen for implementation. Given O&M only provides a basic framework, evolving MOLES required a specialisation of O&M to expand its scope to support structured information. The resulting profile itself went through a number of iterations, resulting in MOLES3.4.

The key specialisations and their similarity to MOLES2 are depicted in Figure 2, which also demonstrates some key changes too. One important element is the shift to a finer granularity in the connection to the archived resource via Observations, though larger collections, represented by Observation Collections, are maintained in line with

the MOLES2 Data Entity level. This explicit refinement over ISO19156 (where it is implicit) allows related Observations to exist, but also supports aggregated datasets within a collection utilising Observations collected for entirely different reasons (i.e. linked to different Projects records). The shift in granularity was an essential element allowing the resulting catalogue to deliver fixed landing pages for dataset DOIs within wider, continually growing collections.

The other major element was the introduction of a much refined way of capturing the process of data production. Details of this are beyond this paper, but it had important implications during migration from MOLES2 to MOLES3 as discussed below.

A full description of MOLES3 is available as a technical note (Ventouras & Lawrence, 2013), with a formal paper in preparation. Although not germane for the material discussed here, it is worth noting there is a strong relationship between the aims and implementations of MOLES3 and W3C PROV (Missier & Moreau, 2013).

Implementing MOLES3 as CEDA-MOLES

Having developed a basic information content model, exploitation involves five further phases: (1) extending the content model to support operational tasks such as editorial control, (2) implementing the appropriate databases, and (3) developing tools to interact with those databases. Next, the resulting databases need to be (4) populated, and (5) the resulting service integrated into organisational procedures.

Extending the Content Model for Operational Tasks

Following the OAIS reference model for open archives (CCSDS, 2012), a key component of the metadata environment is the management and auditing function, ensuring that all parts of the archive are correctly curated and reviewed. In the MOLES2 environment these functions were held in an independent database, leading to inconsistencies between the actual state of the metadata, and the recording of the workflow around that state. Given the focus of MOLES3 to match metadata record granularity to the archive content it describes, a natural structure resulted within which to incorporate these workflow and curation management functions. (This led to another layer of components within the content model, which further compounded the initial implementation problems described below.)

Database Implementation and Tooling

MOLES2 had been implemented using specialisations of Atom documents (Nottingham & Sayre, 2005) stored in an Exist XML database. For MOLES3, obvious implementation choices included continuing the XML document approach or moving to a relational database implementation. In a data management environment there are obvious advantages to both approaches. Document orientation allows a more natural version control of information, whilst relational implementation delivers greater content reusability. Arguably document orientation also allows more flexible content standard migration. As discussed above, however, the practical experience with MOLES2 was that usability issues prevented any exploitation of any theoretical advantages of

documentation orientation, further compounded by the lack of easily usable software tools to build interfaces – too much bespoke framework code was necessary around our Exist implementation. Hence a decision to use a relational implementation was taken.

The initial approach within CEDA sought to exploit a full model-to-database implementation following the "NewMoon" approach (Nagni & Ventouras, 2013), and significant work went into a version which supported a full instantiation of the UML model in a relational database with accompanying Python classes representing all the UML components (Nagni et al., 2012).

The expected advantage of this approach was to allow evolution of the UML model to be realised with greater ease: the underlying database would be able to evolve in line with the model, and code to manipulate the tables would be auto generated. This evolutionary aspect was initially found to be useful, as an iterative approach enabled the operational aspects of the model to be further refined. This approach, however, quickly met some limitations that eventually led to this methodology being abandoned.

The main implementation issue with the underlying databases was caused by the limitations of implicit inheritance and hierarchy within the UML model. In practice, the model was open to circular references, and so additional many-to-many tables were required to resolve such relationships. Nagni and Ventouras (2013) presented a methodology addressing these issues, but this lead to a complex and vast database, spanning 680 tables, of which only 50 were active, containing around 500,000 records each.

Such complexity limited implementation options and performance, preventing the development of sensible, manageable administrative and user interfaces, particularly ones that exploited any of the software frameworks for which CEDA had existing expertise. Dealing with this would have likely resulted in another bespoke software implementation, just like the one with Exist which had been rejected, failing to meet a key requirement of delivering an easily maintainable system that could be further developed from both an administrator/editor and end-user interface perspective.

Consequentially a new approach was adopted, setting the need to use the Django framework[2] as the starting point. Django is a web framework which provides a model/view/controller (MVC) environment with sophisticated database management tools including an object-relational mapper, and it was in significant use in other components of CEDA infrastructure.

To utilise the full Django framework the base UML model was re-examined to establish a simplified profile that could be built entirely within the Django framework. The first stage was to develop a streamlined profile of the model by dropping those whole classes, and class attributes from retained classes, which would be difficult to populate, were unlikely to be used, and were not required operationally (e.g. to support archive linkage, discovery export). Next, class inheritance was avoided by collapsing many classes down, resulting in a much flatter model – one such collapse, for the Observation classes, is shown in Figure 3. As a result a much leaner and cleanly defined model was produced.

Free of the constraints of a pure implementation of a UML model, further refinements were possible within the Django database. Common collections of attributes were identifiable in the main classes and these were replaced with their own class types, but were not implemented in such a way as to reintroduce complexity. These modifications

[2] http://www.djangoproject.com/

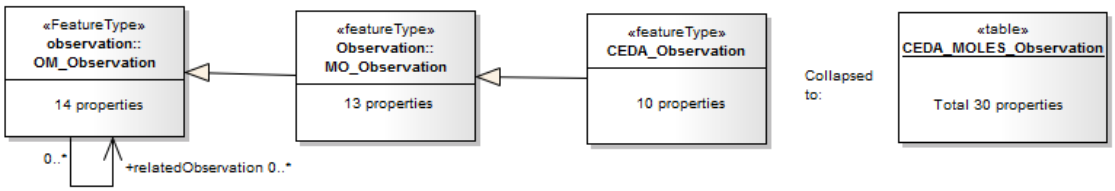

Figure 3. Observation class inheritance (with numbers of properties per class indicated) in the full model and final, flattened model class within Django. Subsequently, the "Observation" class within the database contains most (not all) the inherited properties from the CEDA, MOLES and O&M Observation classes.

were a pragmatic contribution to easing future development and maintenance of the database and surrounding tools and services.

One powerful aspect of using the Django framework was the ability to quickly generate standard user views through the use of Django's templating approach, allowing the focus to remain on crafting the user interface, style and layout of the catalogue largely independent of the content. This exploitation of the MVC paradigm immediately removed backend complexity in that many (particularly user-focused) fields could be constructed, rather than stored, addressing construction issues with new views in MOLES2. Overall maintainability has been, and will be, significantly easier to deliver. One such example is the citation string: within MOLES2 these were all hand crafted and stored by data management staff, and often differed in form between entities and even from content in other fields on the same record. These are now generated from content stored in other fields. Thus, the citation's components – for example the title, author list and URL – will always remain consistent with those on the record itself, and consistent in form between entities.

An unexpected benefit of Django was the support for relatively easy evolution in the database schema and retention of developed tools. The "South"[3] extension to Django was adopted allowing some migration of the database schema and contents so that the data model could be updated without significant disruption to the application and database. This mechanism was of significant use in adapting the model during migration, and it is now clear that improved agility in the infrastructure improves the manageability of the information.

The Migration System

With nearly six thousand records in MOLES2, it would not have been practical to construct MOLES3 records *ab initio*, nor would it have been desirable: traceability of entities to support existing citation was important. Additionally, as MOLES2 was the only source of content such as provenance information, preservation of MOLES2 content remained key to avoiding possible issues with archive migration – a core function of any OAIS-based archive – as highlighted in the migration detailed by (Sawyer et al., 2005). Hence, by design, MOLES3 implemented O&M in such a way as to preserve many aspects of MOLES2 (see Figure 4 for key details). It was thus possible to develop a migration system which preserved and automated as much as was practicable.

[3] http://south.aeracode.org/django_south

 doi:10.2218/ijdc.v10i1.365

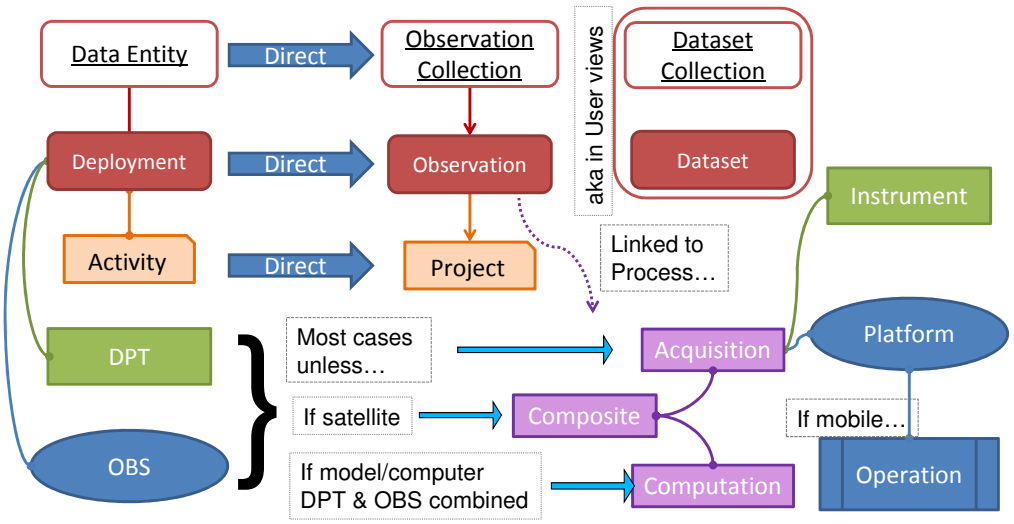

Figure 4. Most MOLES2 objects mapped directly to MOLES3 counterparts, but migration of the Data Production Tools and Observation Stations to the correct Process object(s) was reliant on their sub-type(s) of the MOLES2 objects.

Nevertheless, problems arose in various guises during the migration process: some fields simply lacked any equivalent source content within MOLES2, whilst others could not be mapped due to data type incompatibility. In other cases there was a lack of detail in the MOLES2 content model, or simply missing content in MOLES2 entity (it has already been noted that the lack of integration between management and content functions led to difficulties). Where automation could still help, pragmatic solutions were crafted to resolve these issues. For example, mapping a Data Entity lacking any connected Deployments to an Observation Collection resulted in the creation of a dummy Observation record to hold migrated attributes; whilst sparse Observation records were populated by content inherited from a parent Observation Collection. In other cases, significant human intervention was required, often by amending MOLES2 records so they could be successfully parsed by subsequent migration system runs.

Limitations and Future Expansions

Although the CEDA-MOLES implementation of MOLES3 has only been operational for three months, it has already proven to be very successful, although limitations such as the following have become apparent:

- The content model does not have an adequate mechanism for describing data quality, and it does not constrain relationships sufficiently (e.g. related observations may complement/continue/supersede each other).
- The archive content has heterogeneity that cannot yet be sufficiently supported in the existing content model: non-geospatial environmental digital data (e.g. laboratory data), physical archives (e.g. glass plates), and data with a feature of interest beyond Earth, such as the near-Earth environment and the Martian atmosphere.

Table 1. An indication of the scale of the migration of MOLES2 records to MOLES3 counterparts (upper part). Entities in the lower part of the table were entirely new, although mostly constructed from material inherent in the MOLES2 content. Nearly 1400 unique parties were manually constructed from very poorly controlled party information in MOLES2, and variously connected resulting in nearly 44,000 responsible party links.

MOLES2 component	Number of records	MOLES3 counterpart	Number of records
Data Entity	310	Observation Collection	314
Deployments	3026	Observation	3052
Activity	914	Project	915
Observation Stations	553	Platform	507
Data Production Tools	1012	Instrument	865
		Computation	337
Total MOLES2 records	5815	Total MOLES3 counterpart records	5990

New MOLES3 record types	Number of records
Acquisitions	2594
Composite Processes	245
Party	1397
Responsible Party Info	43754

These two examples demonstrate unplanned issues which have arisen from practice with the expanded capabilities of MOLES3, and planned issues, in that CEDA-MOLES was deliberately restricted to address core issues. The edge cases which remain are primarily related to extra activities accrued beyond core disciplinary activity.

The migration process also demonstrated the consequence of an archive's organic growth over nearly twenty years, in part from changes in both information available, and expectations for how it could and should be used. Ideally much of the MOLES content would be harvested from the archive itself. While some datasets have little or no quality internal archive metadata, it is estimated (Conway et al., 2013) that 41% of files hold useful metadata which could be easily harvested. In particular, where standardised formats and controlled vocabularies have been utilised, such as the CF conventions,[4] such archive metadata would considerably improve the documentation function of MOLES. Hence, the next major investment will be in developing improved capabilities to harvest MOLES relevant material from file-level (archive) metadata. This will be further complemented by enhanced commentary metadata utilising the CHARMe methodology (Blower et al., 2014), to deliver more complete metadata coverage.

On a longer timescale, initial thinking for MOLES4 has already begun. Although the issues above will be addressed and it will be desirable to directly support the W3C PROV content model (Missier & Moreau, 2013), it is likely that the major changes will be in the interface, not the internal structure, exploiting linked data principles to give faceted, and deeper, linked search functionality.

[4] http://cfconventions.org/

Summary

With growing data holdings, and new requirements on content to support new discovery services such as ISO19115 compliance, and data publication, the existing MOLES2 metadata system at CEDA has been upgraded to an implementation of MOLES3.

Utilising ISO19156 Observations and Measurements, MOLES3 has been designed to improve documentation for data holdings, improve navigation between data holdings, and improve the quality of information exported to downstream services. The CEDA-MOLES implementation has also been designed to improve maintainability, both of the infrastructure itself (database schemas and user interfaces) and of the content (via improved information management functions).

The implementation design and the migration of data from MOLES2 to MOLES3 involved important decisions. The MOLES2 implementation was primarily object orientated, being delivered using an XML database, and the MOLES3 implementation is primarily relational, being delivered using a relational database with a modern MVC framework providing object orientated views. In part the decision to move to relational structures was made easier by the vastly improved tooling, such as the Django web framework we utilised, for designing and implementing flexible tools for querying, viewing, and evolving the content. Other obvious immediate benefits of relational structures included that subsidiary entities such as parties would be implemented in a fully relational manner, allowing significant reduction in inconsistencies between records. Although there were major benefits from this approach, however, we have lost the direct connection between the content model (defined in UML) and the implementation (by utilising a manually constructed Django schema in preference to a NewMoon like system), but we believe this to be a relatively small sacrifice. Aspects of the design also reflected the necessity of supporting practical information migration between versions.

Clearly the MOLES2 to MOLES3 migration involved significant intellectual effort (in the content model) and significant development effort (in the systems), but it has also involved a massive investment in time by data scientists. With many thousands of entities auto-migrated, but also thousands manually constructed by hand-parsing original content into formal structures, it demonstrated the importance of ensuring that evolution of standards respect the necessity of maximising entity reuse, both to make migration possible, and to ensure external citation integrity. It also demonstrated that structured information is easier to manage than ad-hoc markup (itself easier to manage than free text). Managing information which includes internal structure held in free text resources is not practical at scale, indeed, despite the major effort in quality control in the migration process, large parts of our new MOLES content will still need human intervention to reach acceptable levels of quality – but this takes time!

References

Blower, J. D., Alegre, R., Bennett, V. L., Clifford, D. J., Kershaw, P. J., Lawrence, B. N., . . . Phipps, R. A. (2014). Understanding climate data through commentary metadata: The CHARMe Project. In Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi & J. Schirrwagen (Eds.), *Communications in Computer and*

Information Science: Vol. 416. Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops (pp. 28–39). doi:10.1007/978-3-319-08425-1_4

Callaghan, S., Lowry, R. & Walton, D. (2012). Data Citation and Publication by NERC's Environmental Data Centres. *Ariadne*, *68*. Retrieved from http://www.ariadne.ac.uk/issue68/callaghan-et-al

CCSDS. (2012). *Reference Model for an Open Archival Information System (OAIS). Issue 2.* (Magenta Book No. CCSDS 650.0-M-2). Washington, DC: Author. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Conway, E., Pepler, S., Parton, G., Garland, W., Newey, C., Jones, H. & Pinder, R. (2013). *The Centre for Environmental Data Archival format audit, appraisal and strategy development.* Paper presented at PV 2013, ESA-ESRIN, Frascati. Retrieved from http://purl.org/net/epubs/manifestation/10941180

International Organization for Standardization. (2013). *Geographic information – Observations and measurements* (Standard No. ISO 19156:2013). Geneva, Switzerland: Author.

Lawrence, B., Lowry, R., Miller, P., Snaith, H. & Woolf, A. (2009). Information in environmental data grids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1890), 1003–1014. doi:10.1098/rsta.2008.0237

Missier, P. & Moreau, L. (2013, April 30). *PROV-DM: The PROV Data Model* (W3C Recommendation). W3C. Retrieved from http://www.w3.org/TR/2013/REC-prov-dm-20130430/

Nagni, M. & Ventouras, S. (2013). Implementation of UML schema in relational databases: A case of geographic information. *International Journal of Distributed Systems and Technologies*, *4*(4), 50–60. doi:10.4018/ijdst.2013100105

Nagni, M., Ventouras, S. & Parton, G. (2012). Implementation of UML schema to RDBM. *Geophyical Research Abstracts*, *14*, 5872. Retrieved from http://meetingorganizer.copernicus.org/EGU2012/EGU2012-5872.pdf

Nottingham, M. & Sayre, R. (2005). *The Atom syndication format* (Proposed Standard No. RFC4287). Fremont, CA: Internet Engineering Task Force. Retrieved from https://tools.ietf.org/html/rfc4287

Sawyer, D., Hills, H. K., McCaslin, P. & Garrett, J. (2005). *Performing a migration in the framework of the OAIS Reference Model: NSSDC case study.* Paper presented at PV 2005, Edinburgh. Retrieved from http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/042.pdf

Ventouras, S. & Lawrence, B. (2013). *Metadata Objects for Linking Environmental Sciences (MOLES): Version 3.4 users guide* (Tech. Rep. No. RAL-TR-2013-001). STFC. Retrieved from http://purl.org/net/epubs/work/65066